Accepted version of the chapter in Wagner, E. and Ockey, G. (eds.) *Emerging Issues in the Assessment of Second Language Listening,* John Benjamins.

Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests

Fumiyo Nakatsuhara, *CRELLA Research Institute, University of Bedfordshire*

ABSTRACT

Examiners intervene in second language oral interviews in order to elicit intended language functions, to probe a candidate's proficiency level or to keep the interaction going. Interventions of this kind can affect the candidate's output language and score, since the candidate is obliged to process them as a listener and respond to them as a speaker.

This chapter reports on a study that examined forty audio-recorded interviews of the oral test of a major European examination board, with a view to examining examiner interventions (i.e., questions, comments) in relation to the listening demands they make upon candidates. Half of the interviews involved candidates who scored highly on the test while the other half featured low-scoring candidates. This enabled a comparison of the language and behaviour of the same examiner across candidate proficiency levels, to see how they were modified in response to the communicative competence of the candidate. The recordings were transcribed and analyzed with regard to a) types

of examiner intervention in terms of linguistic and pragmatic features and b) the extent to which the interventions varied in response to the proficiency level of the candidate.

The study provides a new insight into examiner-examinee interactions, by identifying how examiners are differentiating listening demands according to the task types and the perceived proficiency level of the candidate. It offers several implications about the ways in which examiner interventions engage candidates' listening skills, and the ways in which listening skills can be more validly and reliably measured when using a format based on examiner-candidate interaction.

INTRODUCTION

How examiners and candidates interact during an oral interview test is an area of particular interest to language test developers. Since candidates are required to respond to examiners' questions and interact with them, these tests inevitably require a degree of listening proficiency (e.g., Nakatsuhara, 2012; Seedhouse & Egbert, 2006; see Gary Ockey's chapter for more details). Although it appears obvious that the level of language and the informational density and complexity of examiner

interventions[1] must influence the candidate's ability, as a listener, to respond appropriately, little empirical research has been carried out on the actual listening demands that examiner interventions make upon candidates.

As a first step to understanding listening demands during speaking interview tests, the present study analyzed examiner interventions in the oral test of a major European examination board, the Graded Examinations in Spoken English (GESE) examinations (Trinity College London). This study focused on Grade 7 of the GESE, which targets the B2 level of the Common European Framework of References (Council of Europe, 2001; see Papageorgiou, 2007, for the GESE-CEFR linking study).

The GESE exams aim to assess both speaking and listening skills through face-to-face communicative interaction by "replicat[ing] real-life exchanges in which the candidate and the examiner pass on information, share ideas and opinions and debate topical issues" (Trinity College London, 2009, p.6). A holistic rating scale is currently used, although the use of analytical scales including an 'interactive listening' category has been considered to assess listening more systematically and more explicitly (e.g., Field, 2013a; Inoue & Nakatsuhara, 2014; Nakatsuhara

[1] In this research, the term 'examiner intervention' is used to represent all types of spoken input offered by examiners during an oral interview test, including orally-given task instructions, questions, and responses to the candidate's utterances.

2013). In the current holistic scale, most performance descriptors relate to speaking ability but some also tap into the listening ability required in interaction (e.g., *'The interaction proceeds smoothly, with the candidate contributing promptly and fluently'*).

Additionally, unlike most standard interview tests where all turns are instituted by the examiners' scripted short questions, the GESE examiners do not follow a strict interlocutor framework. Instead, the examiners are instructed to produce a test plan for 'natural interventions' that meet the language requirements of the grade based on the syllabus. Thus, examiners have more freedom and discretion in relation to how they intervene in the interactions. It is desirable to elicit more authentic communication between an examiner and a candidate; however, if examiner interventions are unsystematically varied because of differences in listening demands made on candidates, this could be a potential threat to the validity and reliability of the test.

While the difficult balance between interviewer standardization and the authenticity of test interaction has long been discussed in relation to discourse features between examiners and candidates (e.g., Brown, 2003; Lazaraton, 2002; Ross & Berwick, 1992), the issue has not, until recently, attracted research attention concerning the listening demands for candidates. Thus, this study aimed to offer insights into the listening demands created by examiner interventions in a face-to-face spoken test.

Research questions

This chapter reports on a part of a larger-scale project[2] and addresses two research questions.

RQ1: What types of intervention are used by examiners in a B2-level spoken interview test in terms of their linguistic and pragmatic features?

RQ2: Do the listening demands of examiner interventions differ in relation to the proficiency level of the candidates?

METHODOLOGY

Participants

Audio-recordings of 20 interviews with candidates at Grade A (Distinction) and of 20 interviews with candidates at Grade C (Pass) in Grade 7 were made available by Trinity. In GESE, candidates' performances are rated separately on each of the three phases of the test (see Table 2 below), but the candidates featured were those who received an A or a C on all three phases. To compare interviewer language and behaviour across proficiency levels, samples were selected from

[2] This chapter is based on the larger research project, *'A study of examiner interventions in relation to the listening demands they make on candidates in the GESE exams'* (Nakatsuhara & Field, 2012), funded by Trinity College London.

the same 20 interviewers examining at the two different levels. Steps were taken to ensure that candidates who were paired for the purposes of comparison were as similar as possible in terms of their first language, age and gender, in order to reduce potential effects of test-taker variables. Demographic information about the candidates is summarized in Table 1.

Table 1: Demographic information of the 40 candidates

	First language	**Age**	**Gender**
Grade A candidates	Italian: N=10	Mean=19.45	Male: N=12
(N=20)	Spanish: N=8	Min=10	Female: N=8
	Indian: N=2	Max=39	
Grade C candidates	Italian: N=12	Mean=21.05	Male: N=7
(N=20)	Spanish: N=7	Min=13	Female: N=13
	Indian: N=1	Max=46	

The test

As shown in Table 2, GESE Grade 7 consists of three phases which involve different styles of interaction between the examiner and candidate.

Table 2: Three phases in Grade 7

Phase		**Time**
1	Candidate-led discussion of a topic prepared by the candidate	Up to 5 minutes
2	Interactive task	Up to 4 minutes
3	Conversation on two subject areas selected by the examiner	Up to 5 minutes

In Phase 1 (Topic), the candidate is asked to lead a discussion with the examiner based on a topic the candidate has prepared. Phase 2 (Interactive) requires the candidate to initiate and maintain the interaction by asking the examiner questions, understanding the examiner's responses, and incorporating these into the ongoing discourse to develop the interaction further. Phase 3 (Conversation) involves discussions on two subjects between the candidate and the examiner.

Thus, the topics discussed in Phase 1 are different for each candidate. Examiners are given a choice of prompts for Phase 2 and a choice of topics for Phase 3. The audio-recordings examined in this study included nine prompts in Phase 2 and six topics in Phase 3. While each of the nine prompts in Phase 2 was used the same number of times with Grade A and C cohorts, it was not possible to fully control for the Phase 3 topic variable.

Data analysis

This research draws on Weir's (2005) *socio-cognitive framework for validating language tests* (further elaborated in Geranpayeh & Taylor, 2013, for listening test validation). This instrument was developed to help incorporate the social, cognitive and evaluative (scoring) dimensions of language use systematically into test development and validation (O'Sullivan & Weir, 2011). Its usefulness and practicality have been widely recognised in relation to a number of international language

examinations (e.g., Cambridge General English examinations, the British Council's Aptis test). The socio-cognitive framework as applied to listening was chosen to shape the research design, because it includes a detailed list of contextual parameters that can influence task demands related to the input language, such as nature of information, length, lexical and structural features and speech rate. The framework similarly allows discussion of the cognitive demands associated with the same parameters.

The analysis of the recorded data followed two stages.

Stage 1: All audio-recordings were transcribed using a simplified version of Conversation Analysis (CA) notation (Atkinson & Heritage, 1984).

Stage 2: Examiner interventions were separated from candidates' utterances and then analyzed for the properties shown in Table 3, which correspond to aspects of context validity laid down in the socio-cognitive framework for validating listening tests[3].

Table 3: Analytic measures for examiner interventions

	Contextual parameters	**Measure(s) for the selected parameters**
a	*Syntactic complexity*	Number of verb elements per AS unit
b	*Informational density*	Lexical density
c	*Number and length*	Number and length of interventions
d	*Articulation rate and pauses*	Articulation rate; Number of intra-intervention pauses; Total pause duration [Only for the initial prompting interventions in Phase 2 where examiners read aloud a scripted prompt[4]]

[3] More contextual parameters and more analytic measures to quantify each parameter were examined in Nakatsuhara and Field (2012). However, due to space limitations, this chapter reports only on the parameters and analytic measures presented in Table 3.

[4] This analysis was applied only to the first prompting intervention in Phase 2, mainly due to the labour-intensive nature of the analysis.

The right-hand column of Table 3 shows analytic measures selected to quantify the contextual parameters in the left-hand column. The analytic measures selected relate to observable linguistic and discoursal correlates of these contextual parameters, although there is no assumption that the chosen measures fully cover each of the parameters. Furthermore, it is important to note that there is no preconception as to which of the parameters (if any) is the most critical to the successful communication of meaning, since listening comprehension is a highly integrated process (Field, 2008, p.336-339).

Using these analytic measures, examiner interventions were analyzed to explore the type and variation of their interventions across the three phases of the test (*RQ1*). How each of the selected properties was measured is described in the results and discussion section below. Additional attention was also paid to the extent to which interventions varied between individual examiners. The characteristics of examiner interventions offered to candidates (later) graded A were then compared with those offered to Grade C candidates (*RQ2*). In other words, the analysis compared the level of listening demands created by the examiner input as between stronger and weaker candidates. The goal was to examine whether examiners frame their interventions differently according to their perceptions of candidate proficiency, that is, whether the difficulty of listening input resulting from

examiner interventions was adapted to candidates' ability (as perceived by examiners) or whether listening difficulty was uniform regardless of candidates' proficiency.

In comparing examiner interventions across the three phases of the test and between those provided to Grade A and C candidates, non-parametric tests were used because of the small sample size of the study. Friedman's tests were employed for the analyses of examiner interventions across the three phases, followed by post-hoc comparisons by Wilcoxon Signed Ranks tests (*RQ1*). Mann-Whitney U tests were used for comparing the two groups of candidates (*RQ2*). The value of statistical significance used was $p<0.05$, except for post-hoc comparisons with $p<0.017$ which was the resulting p-value after Bonferroni adjustments to control for multiple comparisons.

The purposes of these analyses are, therefore, to gain some indication as to how the selected contextual parameters are realized by examiners, to what extent they differ from one examiner to another, and whether or not examiners vary their input according to the proficiency level of the candidate.

RESULTS AND DISCUSSION

This section covers (i) examiner interventions across phases, (ii) variation in interventions between examiners, and (iii) variation within examiners in relation to proficiency level.

Syntactic complexity of interventions

The concept of 'syntax' is not always compatible with the analysis of spoken language, since spoken utterances "may consist of single words, phrases, clauses and clause combinations spoken in context" (Carter & McCarthy, 2006, p.167). Nevertheless, the syntactic complexity of input language does have an important role in listening comprehension, as analyzing utterances which consist of more than one clause or entail complex inter-clause relationships potentially impose greater parsing demands upon the listener (Field, 2013b).

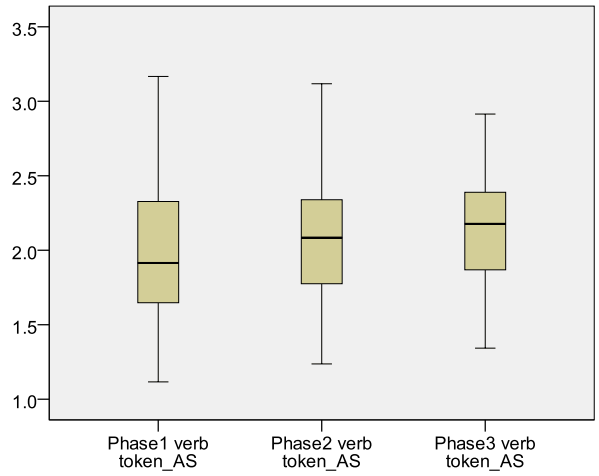
The syntactic complexity of interventions was measured by the number of all verb elements (including to-infinitive and gerund verb forms) divided by the number of AS-units (Foster, Tonkyn, & Wigglesworth, 2000). Given the relatively limited use of subordinate clauses in interactive communication, this measure was suggested as more indicative of syntactic complexity in interactional language (Nitta & Nakatsuhara, 2014), compared to more traditional measures such as the number of subordinate clauses divided by the number of AS-units.

For this measurement, the number of all verb elements in each phase was counted using an automated text analysis tool, TextInspector (Bax, 2011). The results are shown in Table 4 and Figure 1.

Table 4: Number of verb elements per AS-unit

Phase	Min	Max	Median	Friedman's test	Post-hoc comparison (Wilcoxon Signed Ranks test)
1	1.12	3.17	1.91	X^2=6.200; p=0.045	Phases 1 & 2: z=-1.29; p=0.197 (n.s.)
2	1.24	3.12	2.08		Phases 2 & 3: z=-0.645; p=0.519 (n.s.)
3	1.34	2.91	2.18		Phases 1 & 3: z=-2.39; p=0.017 (sig.)

Figure 1: Number of verb elements per AS-unit



Although the median values indicated that examiner interventions were more syntactically complex in Phase 3 followed by Phase 2 and Phase 1, the differences were statistically significant only between Phases 1 and 3.

The finding that Phase 3 interventions were the most structurally complex was in accordance with what one would expect based on the GESE Examiners' Handbook (Trinity College London, 2010, p.10). As in many speaking tests, the final phase is designed to push the candidate's performance to its limit. In Phase 3, examiners have greater control in guiding the interaction, and

they also need to elicit as many as possible of the listed functional, grammatical and lexical items that were specified in the test specifications and that have not already been covered in the first two phases.

Nevertheless, as indicated in Figure 1, examiner interventions in all phases exhibited wide variation between examiners. While some examiners on average used more than three verb elements per AS-unit, other examiners rarely used more complex syntactic elements than one main verb per AS-unit.

The measure of syntactic complexity used was then compared in relation to candidate level. As summarized in Table 5, there was no significant difference between the interventions directed at the two groups. In other words, examiners did not simplify the syntax in their utterances to those candidates they perceived as being of lower ability.

Table 5: Comparisons of the number of verb elements per AS-unit between Grade A and C candidates

Phase	Grade	Min	Max	Median	Mann-Whitney U test
1	AAA	1.30	3.17	1.78	U=146.5; z=-1.448; p=0.148 (n.s.)
	CCC	1.12	2.76	2.08	
2	AAA	1.24	2.94	2.10	U=192.5; z=-0.203; p=0.839 (n.s.)
	CCC	1.55	3.12	2.05	
3	AAA	1.34	2.65	2.20	U=196.0; z=-0.108; p=0.914 (n.s.)
	CCC	1.56	2.91	2.12	

Informational density of interventions

Informational density was measured by lexical density, and lexical density was quantified by the proportion of lexical items (i.e., content words) to total tokens (Halliday, 1985). This measure is

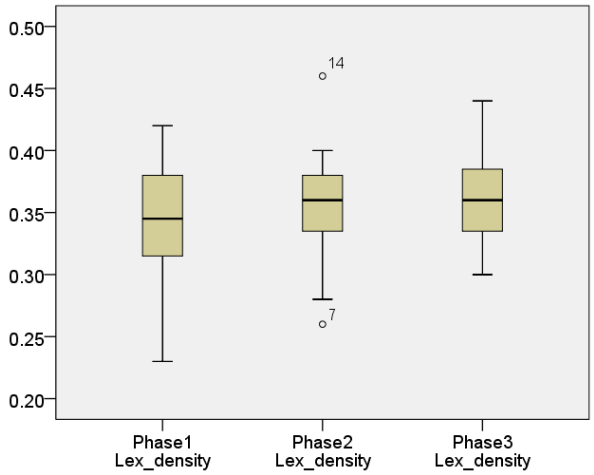
widely used to examine the difficulty of listening texts (e.g., Brunfaut & Révész, 2013) in the recognition that, as the ratio of content words to total tokens increase, there will be denser information in utterances, making it more demanding to process the utterances. Lexical density for each phase session was calculated using TextInspector (Bax, 2011).

Table 6 and Figure 2 indicate that the degree of information density was almost identical across the three phases, and was consistent between examiners. Lexical density is highly dependent upon modality (written vs spoken), and it is commonly considered that a large majority of English spoken texts have a lexical density of under 0.4 (e.g., Ure, 1971), suggesting that information contained in examiner interventions in this test is not unnecessarily dense.

Table 6: Lexical density

Phase	Min	Max	Median	Friedman's test
1	0.23	0.42	0.35	X^2=2.47;
2	0.26	0.46	0.36	p=0.29 (n.s.)
3	0.30	0.44	0.36	

Figure 2: Lexical density



The lexical density of the examiner interventions did not seem to vary according to the level of the candidate. As shown in Table 7, while median values were slightly larger with Grade A candidates in Phases 1 and 2 than with Grade C candidates (as highlighted in the table), they were not significantly different. In other words, both stronger and weaker candidates received similar examiner input in terms of lexical density.

Table 7: Comparisons of lexical density between Grade A and C candidates

Phase	Grade	Min	Max	Median	Mann-Whitney U test
1	AAA	0.23	0.42	0.37	U=139.5; z=-1.643; p=0.100 (n.s.)
	CCC	0.28	0.40	0.34	
2	AAA	0.26	0.40	0.37	U=162.5; z=-1.021; p=0.307 (n.s.)
	CCC	0.29	0.46	0.36	
3	AAA	0.32	0.43	0.36	U=195.5; z=-0.123; p=0.902 (n.s.)
	CCC	0.30	0.44	0.36	

Number and length of interventions

The relationship between utterance length and listening difficulty is generally agreed to be complex, and research does not always support the intuitive position that the longer the utterance is, the more difficult it is to comprehend (e.g., Bloomfield, Wayland, Blodgett & Linck, 2011). Additionally, as learners at CEFR Level B2 (the target of the given test) exercise a considerable degree of automaticity in lower-level cognitive processes in listening (Field, 2008), they can be assumed to be capable of handling longer utterances. The same consideration no doubt applies to the number of utterances they have to process in a single listening event. Despite this, it has been asserted that the number and

length of examiner interventions do indeed contribute to difficulty, in terms of "the cumulative effect of meaning construction" and the listener's need to hold a developing discourse representation in his/her mind (Elliott & Wilson, 2013, p.195).

On these grounds, the number and length of interventions (i.e., average number of words per intervention) were measured. Table 8 and Figures 3 and 4 indicate that the number of interventions was the largest in Phase 3 followed by Phases 1 and 2, and the difference between Phases 2 and 3 approached significance. For the length of interventions, Phase 2 had the longest interventions followed by Phases 3 and 1, and all pairings in the phases showed a significant difference.

Table 8: Number and length of interventions

Phase	Min	Max	Median	Friedman's test	Post-hoc comparison
Number of interventions					
1	8.00	47.00	17.50	X^2=8.15; p=0.017 (sig.)	Phases 1 & 2: z=-1.674; p=0.094 (n.s.)
2	6.00	33.00	16.50		Phases 2 & 3: z=-2.312; p=0.021 (n.s.)*
3	7.00	41.00	19.50		Phases 1 & 3: z=-0.960; p=0.337 (n.s.)
Length of interventions (average number of words per intervention)					
1	5.30	15.05	9.48	X^2=17.22; p<0.001 (sig.)	Phases 1 & 2: z=-4.032; p<0.001 (sig.)
2	6.82	27.50	12.26		Phases 2 & 3: z=-2.512; p=0.012 (sig.)
3	7.27	30.78	11.30		Phases 1 & 3: z=-3.307; p=0.001 (sig.)

*The Bonferroni adjustment was made to the alpha level (0.05/3=0.017).

Figure 3: Number of interventions

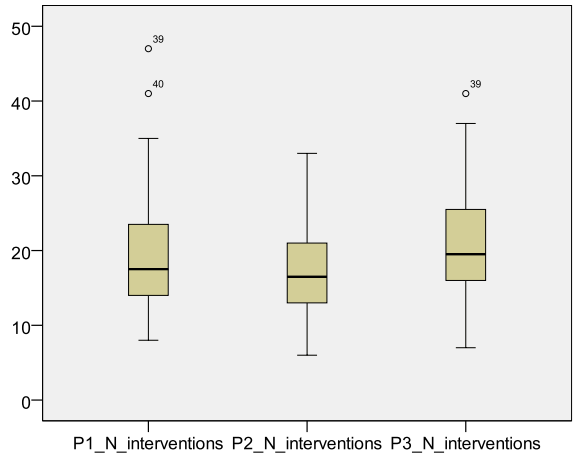
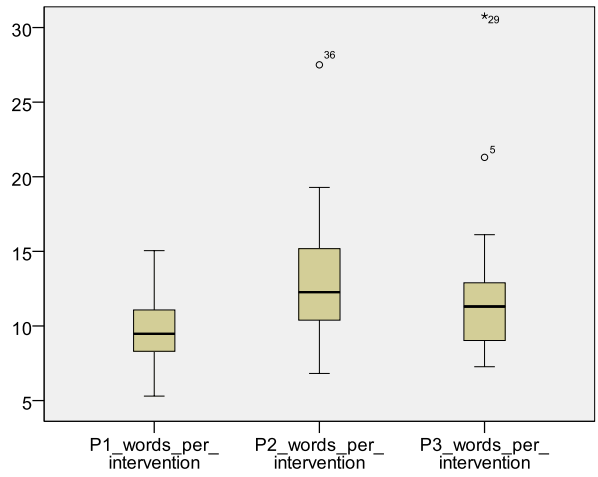


Figure 4: Length of interventions (Number of words per intervention)



All these findings seem to be congruent with the roles of the examiner interlocutor in these three tasks. Phase 3 is the conversation phase, where examiners are required to take a lead in discussing two subject areas, thus causing them to intervene more frequently. By contrast, Phase 2 is the interactive phase, where it is essentially the candidate's responsibility to initiate and maintain the discourse and examiners chiefly intervene when the turn is ceded to them. It also seems logical that Phase 2 interventions tended to be longer than those in Phases 1 and 3, as Phase 2 interventions are

in effect responses to the candidate's questions seeking information, opinions and suggestions. It is similarly understandable that Phase 3 interventions were longer than those in Phase 1. Examiner interventions in Phase 1 mainly serve to facilitate the candidate-led discussion on a topic prepared by the candidate, while examiners select subject areas themselves and guide the conversation in Phase 3.

Despite these general trends in line with the task types, the number and length of interventions varied widely between examiners. For example, in Phase 2, the number of interventions ranged from 6 to 33, and the length of interventions that candidates received could be as short as 7 words or as long as 27 words.

The number and length of interventions directed at Grade A candidates were then compared with those directed at Grade C candidates. As shown in Table 9, examiner interventions with Grade A candidates tended to be, on average, greater in these two measures (an exception was the length of interventions in Phase 1; as highlighted in the table). However, none of these trends were statistically significant and there was considerable variation between examiners (see minimum and maximum values).

Table 9: Comparisons of the number/length of interventions between Grade A and C candidates

Phase	Grade	Min	Max	Median	Mann-Whitney U test
			Number of interventions		
1	AAA	13	47	20.00	U=153.0; z=-1.275; p=0.202 (n.s.)

	CCC	8	41	16.00	
2	AAA	11	33	17.50	U=140.0; z=-1.626; p=0.104 (n.s.)
	CCC	6	32	15.50	
3	AAA	7	41	20.50	U=184.5; z=-0.420; p=0.674 (n.s.)
	CCC	8	37	19.00	
			Length of interventions		
1	AAA	5.30	15.05	9.18	U=142.5; z=-1.555; p=0.120 (n.s.)
	CCC	5.55	14.08	10.54	
2	AAA	6.82	18.92	12.26	U=192.0; z=-0.216; p=0.829 (n.s.)
	CCC	7.59	27.50	12.10	
3	AAA	7.27	21.30	11.72	U=179.0; z=-0.568; p=0.570 (n.s.)
	CCC	7.94	30.78	10.35	

Articulation rate and intra-intervention pauses

In the beginning of Phase 2 (Interactive) only, examiners select one prompt from the list provided and "read aloud the provided rubric clearly and fairly slowly once" (Trinity College London, 2010, p.10). The initial prompt is thus key to the candidate's ability to perform the task successfully. A typical prompt[5] might be, *"It's my friend's birthday soon. I'd like to organize something special, but I'm not sure what to do."*

The analysis was designed to measure the extent to which examiners adopted the required 'clear and slow' speaking style in terms of articulation rate and pauses within a single intervention. It also investigated if there were marked differences in the clarity and speed of prompting in interventions directed towards candidates graded C as against those graded A. Given the online nature

[5] This was taken from Trinity's online sample video resources (http://www.trinitycollege.com/site/?id=3108).

of listening, a faster speech rate is taken to have a detrimental effect on comprehension, although the effect does not seem to be linear (e.g., Griffiths, 1992).

In principle, the speed of speech can be altered either by the articulation rate of the speaker or by the number and length of pauses. However, it has long been demonstrated that pausing plays a greater role in slowing speech than does a reduced rate of articulation (Goldman-Eisler, 1968, p.24). Pauses not only allow the listener more time to process input, but also serve to mark word boundaries more clearly, thus assisting the process of lexical segmentation (Field, 2013b, p.118).

The number of syllables in each prompting intervention was counted using TextInspector (Bax, 2011). Audacity software was then used to measure speech time, the number of pauses over 0.2 seconds, and each pause time above 0.2 seconds. Different researchers use different criteria for determining a pause. Kormos and Denes (2004) counted pauses of more than 0.2 seconds, while more lenient cut-off points were set by other researchers, e.g., 0.25 seconds by De Jong, Steinel, Florijn, Schoonen and Hulstijn (2012) and 0.4 seconds by Tavakoli and Foster (2008). For the present analysis, it was decided to use 0.2 seconds as a cut-off point, as repeated listening of the recordings of the Phase 2 prompts by the researcher and colleagues confirmed that pauses of 0.2 seconds and above seemed recognizable to the listener. Articulation rate was calculated by total number of syllables

divided by total duration of speech. Both filled pause time and unfilled pause time were excluded from the analysis of articulation rate, but were included in the analysis of pausing patterns.

Results are presented in Table 10. The literature on speaking rate indicates that native speakers of English generally articulate around 4.1 syllables per second when reading a passage (Gut, 2007, p.78). On this criterion, the prompts that examiners read aloud in this part of the test were clearly slow. Also, these prompting interventions possessed, on average (by median) two pauses that were over 0.2 seconds and that together totalled 1.15 seconds. Given that the prompting interventions are relatively short (i.e., Total turn time: Median=10.0 seconds), the number of pauses and the pause duration seem to indicate that these interventions were delivered with special care to provide the candidates with the time to process incoming listening input accurately. The phonetics literature (Goldman-Eisler, 1968) suggests that speech perceived as slow is largely characterized by greater length of pausing. The present findings were very much in line with this; the variation between individual examiners was caused more by their pausing patterns than by the speed of articulation (as indicated by the minimum and maximum values as well as the standard deviation of each measure in Table 10).

Table 10: Articulation rate and pauses in Phase 2 prompting interventions

Measure	Min	Max	Median	Mean	SD
Articulation rate	2.36	4.40	3.30	3.30	0.45
Number of pauses	0.00	9.00	2.00	2.55	1.99

Total pause time (in seconds)	0.00	5.40	1.15	1.28	1.13

These prompting interventions were then compared between Grade A and C candidates. Table 11 indicates that none of the measures showed a statistically significant difference in relation to the level of the candidates.

Table 11: Comparisons of articulation rate and pauses between Grade A and C candidates

Measure	Grade	Min	Max	Median	Mann-Whitney U test
Articulation rate	AAA	2.36	4.16	3.35	U=196.0; z=-0.108; p=0.914 (n.s.)
	CCC	2.53	4.40	3.25	
Number of pauses	AAA	0.00	5.00	2.00	U=151.5; z=-1.335; p=0.182 (n.s.)
	CCC	0.00	9.00	3.00	
Total pause time (in seconds)	AAA	0.00	3.50	1.00	U=155.5; z=-1.206; p=0.228 (n.s.)
	CCC	0.00	5.40	1.55	

However, although this was not reflected in the statistical significance, during the analysis of pausing patterns, it was noted that some Grade C candidates received a larger number of pauses within the prompting intervention. When the candidates were listed in the order of the pause frequency and total pausing times received in the prompting interventions, seven Grade C candidates were included in both top ten lists. Therefore, there might have been a tendency for weaker candidates to be given more frequent, longer pauses in their Phase 2 prompting interventions, thus assisting their listening. This could suggest that some examiners were already sensitive to the listening proficiency of a candidate by Phase 2, and slowed down their speech by means of additional pausing when delivering the Phase 2 prompts to lower proficiency candidates.

Language functions used for interventions

It was assumed that, at the B2 level, candidates were likely to have the requisite knowledge of syntax and lexis to be capable of following most or all of the language used in examiner interventions. Where their knowledge might be less complete was in the range of pragmatic functions that examiners chose to express (Council of Europe, 2001). These functions potentially range from simple to complex in terms both of their wording and of the explicitness of the relationship between linguistic form and speaker intention. The ability to identify and interpret the illocutionary intentions of a speaker (Levinson, 1983), realized in different degrees of explicitness, is an important contributory factor in higher-level cognitive processes in listening (Field, 2013b, p.100).

To obtain insights into both the types and explicitness of the language functions used by examiners, an analysis was conducted using a modified version of O'Sullivan, Weir and Saville's (2002) function checklist, which has been successfully applied to interactions in various speaking tests (e.g., Brooks, 2003; O'Sullivan, Taylor & Wall, 2011). The checklist consists of an extensive table of 11 *informational* functions (e.g., *expressing an opinion, justifying an opinion*), 15 *interactional* functions (e.g., *asking for information, negotiating meaning*) and 4 functions which entail *managing interaction* (e.g., *initiating, changing topics*). While it was relatively straightforward

to code the interaction transcripts in terms of these functions, the question of their explicitness required more detailed examinations (taking into account, for example, how each function was encoded and whether its illocutionary force seemed to be understood by the candidate). It should therefore be noted that findings from the latter analysis are at best suggestive. The recorded material analysed related to interactions which had taken place before the study; it was therefore impossible to investigate examiner intentions or listener interpretation more directly.

The use of the three main types of functions are visually illustrated in Figures 5-7. There were differences between the three phases in the ways they engaged the examiner. Examiner interventions in Phase 1 (Topic) were made more for interactional purposes than informational or discourse management purposes. In particular, examiners frequently intervened to *ask for opinions, ask for information,* and *make comments*. In contrast, examiner interventions in Phase 2 (Interactive) were characterised more as informational, such as *giving personal information, expressing opinions* and *describing things, people and events*. The interventions observed in Phase 3 (Conversation) were somewhat similar to those in Phase 1, but there were more varied interactional interventions (e.g., *modifying and expanding the candidate's utterance*). More interventions for interactional management purposes were also observed in Phase 3 than the other two phases.

Figure 5: Language functions used for interventions in Phase 1 (Topic)

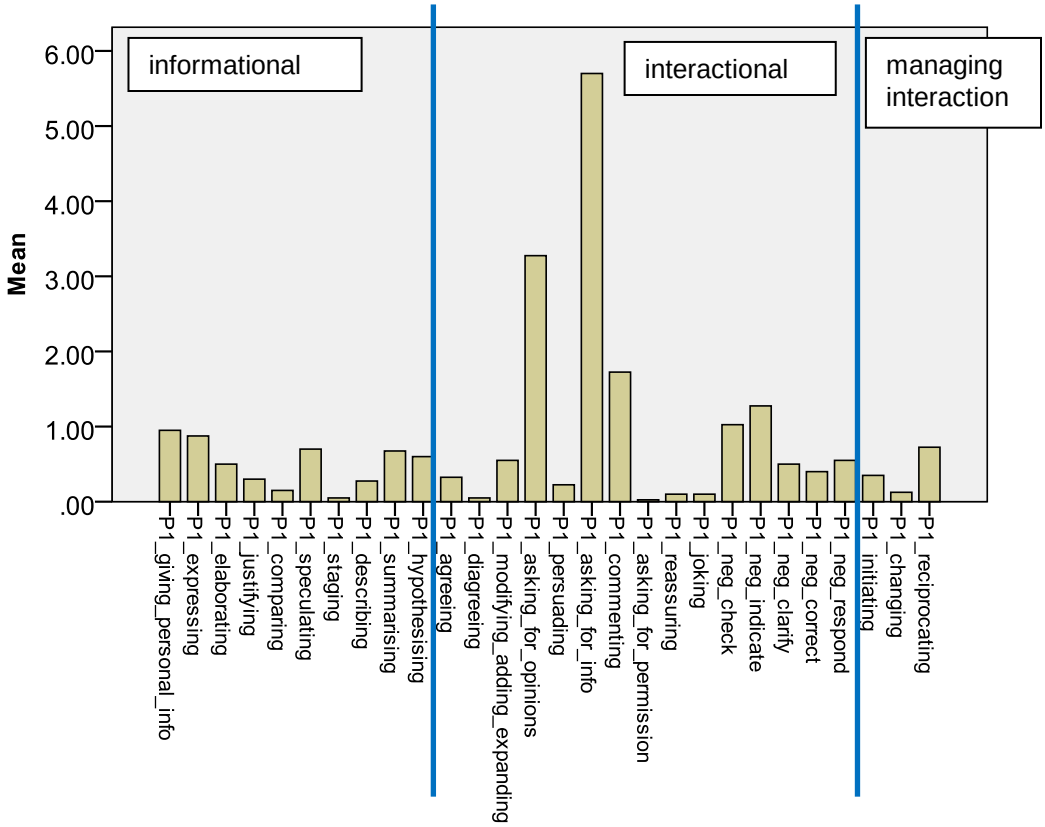


Figure 6: Language functions used for interventions in Phase 2 (Interactive)

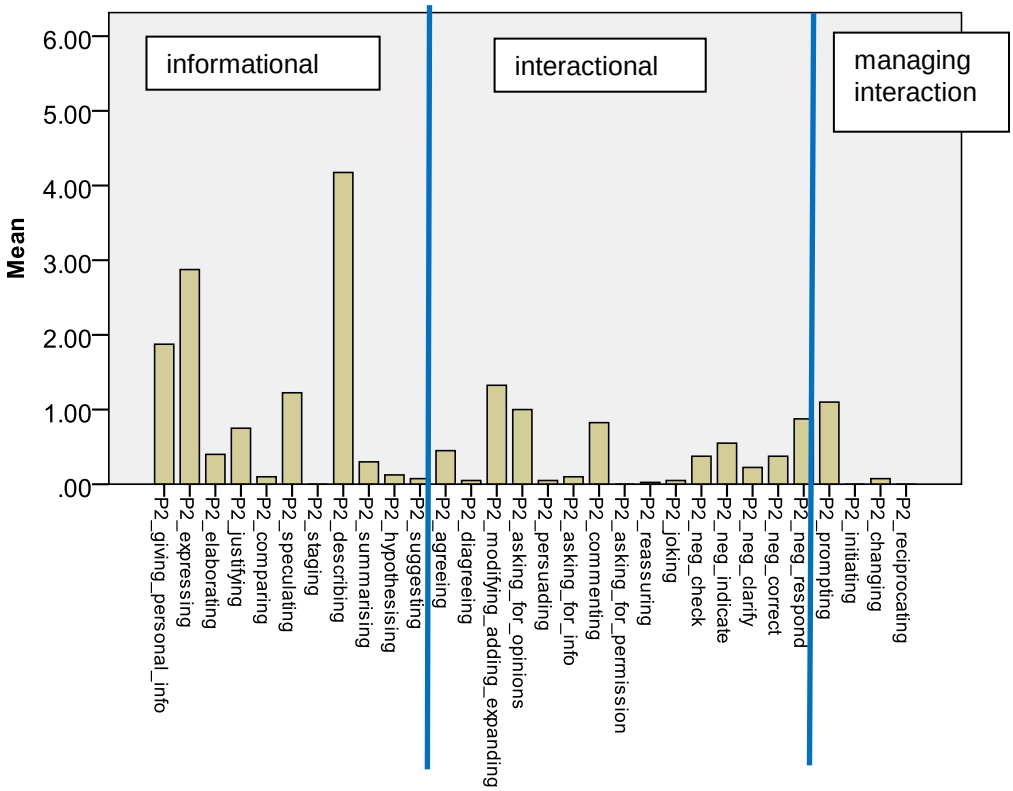
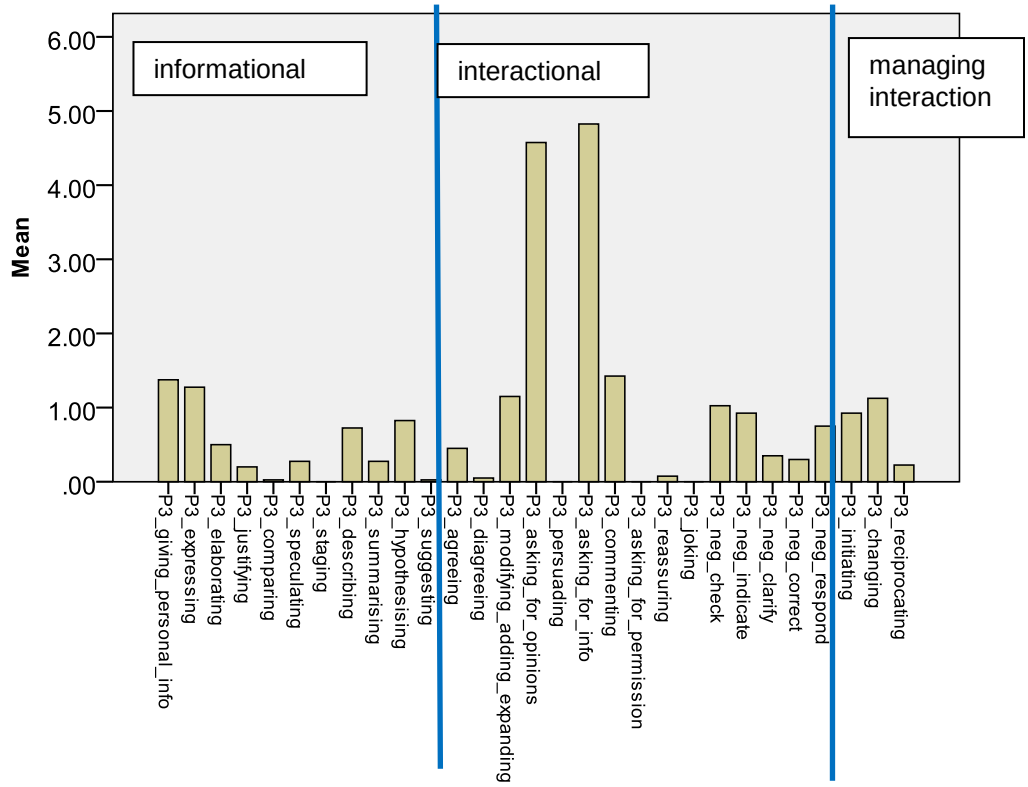


Figure 7: Language functions used for interventions in Phase 3 (Conversation)



In general, the findings suggest that a multi-phase face-to-face test like the one investigated in this study entails interventions which embrace a wide range of purposes. The result is that the listener is faced with a variety of pragmatic functions that need to be interpreted. Examination of the transcripts (see Nakatsuhara & Field, 2012) indicated that these pragmatic functions seemed to vary in terms of the ease with which the candidates were able to interpret them. Certain pragmatic intentions were likely to be consistently clear enough for the candidate listener (e.g., *providing personal opinions*) with the result that no misunderstanding was caused. Others seemed inherently more difficult to interpret due to the less explicit nature of the pragmatic concept (e.g., *hypothesizing*, *modifying*); these sometimes led to misunderstandings or to clarification requests.

Each of the functions used for interventions was then compared between interventions provided for Grade A and C candidates. Table 12 presents the nine functions which showed a significant difference in relation to the proficiency level of the candidates.

Table 12: Comparisons of purposes for interventions between Grade A and C candidates

Function (Phase)	Grade	Min	Max	Median	Mann-Whitney U test
Expressing opinions/preferences (Phase 3)	AAA	0	7	1.50	U=110.0; z=-2.572; p=0.010
	CCC	0	3	0.00	(A>C)
Speculating (Phase 3)	AAA	0	3	0.00	U=129.5; z=-2.627; p=0.009
	CCC	0	1	0.00	(A>C)
Describing (Phase 3)	AAA	0	5	0.50	U=133.5; z=-2.078; p=0.038
	CCC	0	2	0.00	(A>C)
Agreeing (Phase 2)	AAA	0	3	0.00	U=138.0; z=-2.079; p=0.038
	CCC	0	2	0.00	(A>C)
(Phase 3)	AAA	0	4	0.00	U=137.0; z=-2.251; p=0.024
	CCC	0	1	0.00	(A>C)
Commenting (Phase 2)	AAA	0	5	1.00	U=160.0; z=-2.082; p=0.037
	CCC	0	2	0.00	(A>C)
Negotiating meaning: indicating understanding (Phase 1)	AAA	0	6	1.00	U=189.0; z=-1.979; p=0.048
	CCC	0	6	1.00	(A>C)
Asking for information (Phase 3)	AAA	0	14	3.00	U=121.0; z=-2.154; p=0.031
	CCC	0	14	5.00	(C>A)
Negotiating meaning: correcting an utterance made by other speaker (Phase 3)	AAA	0	1	0.00	U=119.0; z=-2.901; p=0.004
	CCC	0	2	0.00	(C>A)
Negotiating meaning: responding to requests for clarification (Phase 1)	AAA	0	2	0.00	U=101.5; z=-3.155; p=0.002
	CCC	0	5	1.00	(C>A)
	AAA	0	2	0.00	U=133.0; z=-2.030; p=0.042
(Phase 3)	CCC	0	4	1.00	(C>A)

It would appear that when examiners expanded upon their interventions with Grade A candidates, their intentions often reflected a desire to participate to a greater degree in the interaction by *expressing their own opinions and preferences, speculating, describing* people/events, *agreeing with* what the candidate said, *commenting on* what they said and *indicating understanding* of what they said. In contrast, the examiner interventions that were more frequent in sessions with Grade C

candidates seemed to be associated with keeping the interaction going by *asking for information*, *repairing conversation by correcting an utterance made by the candidate*, and *responding to a request for clarification*. While some caution must be observed in making a connection between the types of function used and their relative difficulty in terms of listening, the results of this analysis suggest that examiners offered stronger candidates more opportunities to comprehend and respond to a range of language functions, while weaker candidates tended to receive language functions of a scaffolding nature.

CONCLUSION

The goal of this study was to gain a better understanding of the listening demands upon candidates in a multi-phase spoken test involving an examiner who both interviews and interacts with the candidate. With this in mind, it quantified linguistic and pragmatic features of the examiners' language that served as spoken input to the candidates. The findings suggest that interactive speaking tests in this format have the potential to engage the candidate listener to different degrees according to the types of task and the role taken by the examiners as the source of listening input. However, the data also presents a potential threat to the validity and reliability of such tests due to considerable variation between examiners in terms of certain aspects of their interventions. As noted in the introductory

section of this chapter, the issue of striking a balance between interlocutor standardization and authenticity of interaction seems to raise important issues in relation to the listening demands posed upon candidates during interactive spoken assessments.

It was found that some examiners appeared to take account of a candidate's ability level by adjusting their input in terms of informational density, number and length of interventions, pausing patterns in Phase 2 prompting, and the language functions used for interventions. However, not all of these results were statistically significant, suggesting a degree of variation between individual behaviour. In the case of those who did adjust their interventions, it was interesting to observe that these examiners had, quite early on in the test, succeeded in identifying the likely proficiency level of the candidate and responded accordingly.

These findings offer several implications about the ways in which examiner interventions engage candidates' listening skills, and the ways in which listening skills can be more validly and reliably measured when using a format based on examiner-candidate interaction. Examiners need to be made more aware of the part played by listening in the test tasks, trained and standardized in their linguistic and pragmatic features with a view to assessing listening during their interaction. While it might not be practical or possible to formally specify the way in which interventions can be calibrated level by level, examiner training should certainly feature how to test candidates' listening proficiency

by differentiating more clearly between the listening demands of their interventions. Attention should also be drawn to the need to evaluate candidates' listening and speaking skills separately, since a candidate who performs poorly in production may nonetheless achieve a good level of understanding, and vice versa (Nakatsuhara & Field, 2012). If listening is a part of the test construct, it seems essential to make use of analytic rating scales that include a 'listening' category. This should heighten examiners' awareness of the role of listening skills and serve to alert examiners to the fact that a candidate's listening and speaking proficiency might well differ markedly.

By affording a new insight into examiner-candidate interactions, this study has made a first step towards more effective assessment of the essential part played by listening competence in successful L2 communication. It is to be hoped that future monitoring and validation of interactive spoken tests will serve to introduce greater awareness of the skill into the way such tests are delivered.

ACKNOWLEDGEMENTS

I am very grateful to the editors of this book, the two anonymous reviewers and Dr John Field for their insightful comments on an earlier draft of this chapter. This research was funded by Trinity College London, and carried out under the Trinity Funded Research Programme.

REFERENCES

Atkinson, J. M. and Heritage, J. (1984) *Structures of social action.* Cambridge, New York: Cambridge University Press.

Bax, S. (2011) *TextInspector*. Retrieved November 15, 2011 from http://www.textinspector.com/ .

Bloomfield, A.N., Wayland, S. C., Blodgett, A. and Linck, J. (2011) Factors related to passage length: Implications for second language listening comprehension, *CogSci 2001 Proceedings*: 2317-2322

Brooks, L. (2003) Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes, 11,* 20-21.

Brown, A. (2003) Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brunfaut, T. and Révész, A. (2013) Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*(1), 31-65.

Carter, R. A. and McCarthy, M. J. (2006) *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Council of Europe (2001) Common European Framework of Reference for Languages: tearning, teaching, assessment. Cambridge: Cambridge University Press.

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of Speaking Proficiency. Studies in Second Language Acquisition, 34(01), 5–34.

Elliott, M. and Wilson, J. (2013) Context validity. In A. Geranpayeh and L. Taylor (eds.), *Examining Listening: Research and practice in assessing second language listening, Studies in Language Testing vol. 35.* Cambridge: UCLES/Cambridge University Press.

Field, J. (2008) Listening in the languag*e classroom.* Cambridge: Cambridge University Press.

Field, J. (2013a) *The assessment of listening proficiency in Trinity tests of spoken interaction: Guidelines for examiners*. Internal research report, Trinity College London.

Field, J. (2013b) Cognitive validity. In A. Geranpayeh and L. Taylor (eds.), *Examining Listening: Research and practice in assessing second language listening, Studies in Language Testing vol. 35.* Cambridge: UCLES/Cambridge University Press.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000) Measuring spoken language: A unit for all reasons. *Applied Linguistics, 2*(3), 354–375.

Geranpayeh, A. and Taylor, L. (eds.) (2013) *Examining listening: Research and practice in assessing second language listening. Studies in Language Testing vol. 35*. Cambridge: UCLES/Cambridge University Press.

Goldman-Eisler, F. (1968) *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Griffiths, R. (1992) Speech rate and listening comprehension: Further evidence of the relationship, *TESOL Quarterly*, *26*(2): 385-390.

Gut, U. (2007) Foreign accent. In C. Müller (ed.), *Speaker Classification: Fundamentals, features, and methods* (pp. 75-87). Berlin: Springer.

Halliday, M. A. K. (1985) *Spoken and Written Language*. Geelong, Vic.: Deakin University Press.

Inoue, C. and Nakatsuhara, F. (2014) *Trinity College London Integrated Skills in English (ISE): Speaking and Listening - Phase 3 pilot analysis.* Internal research report, Trinity College London.

Kormos, J. and Denes, M. (2004) Exploring measures and perceptions of fluency in the speech of second language learners, *System, 32*(2):45-164.

Lazaraton, A. (2002) *A qualitative approach to the validation of oral language tests. Studies in Language Testing vol. 14.* Cambridge: UCLES/Cambridge University Press.

Levinson, S.C. (1983) Pragmatics. Cambridge: Cambridge University Press.

Nakatsuhara, F. (2012) The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking test. In L. Taylor and C. J. Weir (eds.), *IELTS Collected Papers 2: Research in reading and listening assessment* (pp. 519-573). Cambridge: UCLES/Cambridge University Press.

Nakatsuhara, F. (2013) *Trinity College London Integrated Skills in English (ISE) 'The Interview': Reviewing speaking rating criteria and rating procedures.* Internal research report, Trinity College London.

Nakatsuhara, F. and Field, J. (2012) *A study of examiner interventions in relation to the listening demands they make on candidates in the GESE exams,* Internal research report, Trinity College London.

Nitta, R. and Nakatsuhara, F. (2014) A multifaceted approach to investigating pre-task planning effects on paired oral test performance, *Language Testing*, *31*(2): 147-175.

O'Sullivan, B., Taylor, C., and Wall, D. (2011) *Establishing evidence of construct: A case study.* Paper presented at the 8th annual EALTA conference, Siena, Italy.

O'Sullivan, B. and Weir, C.J. (2011) Language Testing and Validation. In B. O'Sullivan (ed.) *Language Testing: Theory and Practice* (pp. 13–32). Oxford: Palgrave.

O'Sullivan, B., Weir, C.J., and Saville, N. (2002) Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

Papageorgiou, S. (2007) *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference.* London: Trinity College London.

Ross, S. and Berwick, R. (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(2), 159-176.

Seedhouse, P. and Egbert, M. (2006) The interactional organisation of the IELTS Speaking Test. In P. McGovern and S. Walsh (eds.), *IELTS Research Report Volume 6* (pp. 161-205). Canberra: British Council & IDP Australia.

Tavakoli, P. and Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. Language Learning. 58(2): 439-473.

Trinity College London (2009) *Graded Examinations in Spoken English (GESE) Syllabus - from 1 February 2010.* London: Trinity College London.

Trinity College London (2010) *Examiners' Handbook from 2010: strictly confidential - for examiner use only.* London: Trinity College London.

Ure, J. (1971) Lexical density and register differentiation. In J.E. Perren and J.L.M. Trim (eds.), *Applications of linguistics* (pp. 443-452). Cambridge: Cambridge University Press

Weir, C. J. (2005) *Language testing and validation: An evidence-based approach.* London: Palgrave Macmillan.